

Метод наименьших углов для логистической регрессии

К. С. Скипор, В. В. Стрижов

Московский физико-технический институт

Вычислительный центр РАН, Москва

e-mail: skiporkonstantin@mail.ru, strijov@ccas.ru

29.04.2010

Предлагается и исследуется алгоритм отбора признаков для решения задач восстановления логистической регрессии. Алгоритм основан на методе наименьших углов для модели линейной регрессии с использованием дополнительно линеаризации функционала качества. Приводится математическое обоснование предложенного алгоритма. Работа алгоритма проиллюстрирована задачей изучения факторов риска ишемических заболеваний сердца.

1. Введение

В работе рассматривается отыскание из множества признаков такого его подмножества, для которого их линейная комбинация наиболее точно описывает данные. В 1966 году Дрейпером был предложен ступенчатый алгоритм выбора признаков (Forward Stagewise) [1–3].

EXAMPLE

2. Постановка задачи отбора признаков

Дана выборка $D = \{(\mathbf{x}^i, y^i)\}_{i=1}^m$, в которой i -й объект описывается строкой из n числовых признаков, $\mathbf{x}^i = (x_j^i)_{j=1}^n \in \mathbb{R}^n$ и метки класса $y^i \in \{0, 1\}$. Верхний индекс i указывает порядковый номер объекта выборки, нижний индекс j — порядковый номер признака. Векторы признаков $\mathbf{x}_j = (x_j^1, \dots, x_j^i, \dots, x_j^m)^T$ являются линейно независимыми свободными переменными, а вектор $\mathbf{y} = (y^1, \dots, y^i, \dots, y^m)^T$ является зависимой переменной.

EXAMPLE

Требуется построить такой алгоритм последовательного добавления признаков, что на каждом шаге:

- определяются набор *активных признаков* с *активным множеством* индексов \mathcal{A} и соответствующий набору ненулевой вектор параметров $\beta_{\mathcal{A}}$, такой что $\beta_{\mathcal{A}^c} = \mathbf{0}$, $\mathcal{A} \sqcup \mathcal{A}^c = \{1, \dots, n\}$;
- набор *активных признаков* и вектор параметров $\beta_{\mathcal{A}}$ доставляют максимум приращению логарифма правдоподобия ℓ ;
- скорость роста функционала ℓ по любому активному признаку не меньше скорости роста по любому неактивному признаку.

3. Описание алгоритма

Будем считать, что принята линейная модель

$$\mathbf{y} = \boldsymbol{\mu}(X, \boldsymbol{\beta}) + \varepsilon,$$

где функция регрессии $\boldsymbol{\mu}(X, \boldsymbol{\beta})$, представляющая собой приближение вектора \mathbf{y} , имеет вид

$$\boldsymbol{\mu}(X, \boldsymbol{\beta}) = \sum_{j=1}^n \mathbf{x}_j \beta_j = X\boldsymbol{\beta}, \quad (1)$$

Критерием качества назначена среднеквадратичная ошибка

$$S(X, \boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{\mu}(X, \boldsymbol{\beta})\|^2.$$

Замечание 1. Как и в случае *LARS*, вектор *EXAMPLE*.

Перейдем теперь к формальной постановке решаемой задачи.
EXAMPLE

Лемма 1. Задача *EXAMPLE*.

Для доказательства теоремы 1 сформулируем некоторые вспомогательные утверждения.

Лемма 2. Пусть векторы $\mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{a}_{k+1} \in \mathbb{R}^n$ линейно независимы. *EXAMPLE*

Для использования леммы 2 определим понятие аффинной зависимости векторов [4].

Определение 1. Точки $\mathbf{a}_1, \dots, \mathbf{a}_k \in \mathbb{R}^n$ называются аффинно зависимыми, если существуют $\lambda_1, \dots, \lambda_k$, не равные нулю одновременно и такие, что

$$\sum_{i=1}^k \lambda_i \mathbf{a}_i = 0, \quad \sum_{i=1}^k \lambda_i = 0.$$

EXAMPLE

С помощью следующей теоремы формулируется утверждение о решении задачи линейного программирования (1).

Теорема 1. Если ЗЛП (1) имеет решение \mathbf{g}^* , то

$$\mathbf{g}^* \in \Upsilon, \quad (2)$$

причем

$$\mathbf{g}^* \in \Upsilon, \quad (3)$$

где " \min^+ " означает, что минимум берется только из положительных значений.

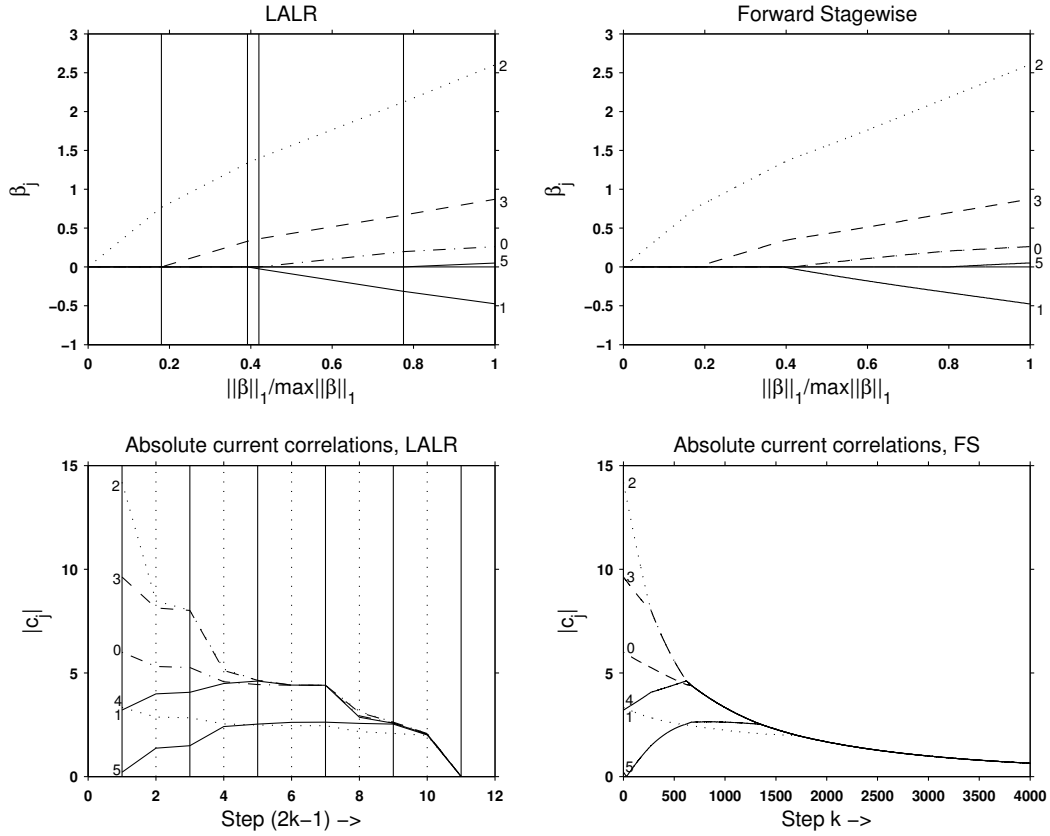


Рис. 1. Сравнение оценок коэффициентов для LALR и Forward Stagewise для модельных данных. Номера кривых соответствуют номерам признаков. Сплошные вертикальные линии обозначают шаги, а штриховые вертикальные — дополнительную итерацию для каждого шага.

4. Вычислительный эксперимент

Сравним предложенный алгоритм с описанным в [2, 3] итеративным алгоритмом Forward Stagewise. На каждом шаге алгоритм выбирает признак \mathbf{x}_{j^*} , имеющий наибольшую корреляцию c_{j^*} с текущим вектором остатков $\mathbf{y} - \mathbf{s}(\mathbf{m})$ и делает небольшое смещение γ текущего приближения в направлении выбранного признака \mathbf{x}_{j^*} ,

$$j^* = \arg \max |c_j| \quad \text{и} \quad \boldsymbol{\mu} \rightarrow \boldsymbol{\mu} + \gamma \text{sign}(c_{j^*}) \mathbf{x}_{j^*}.$$

Чем меньше абсолютная величина смещения γ , тем точнее получается оценка параметров $\boldsymbol{\beta}$.

Сгенерируем $m = 50$ объектов с пятью независимыми, нормально распределенными признаками $\mathbf{x}_1, \dots, \mathbf{x}_5$, т.е. $\mathbf{x}_i = (x_{i1}, \dots, x_{i5}) \sim \mathcal{N}_5(\mathbf{0}, \mathbf{I})$. Примем модель

$$\mathbf{y} = \frac{1}{1 + \exp(-(\beta_0 + \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \mathbf{x}_3\beta_3))} + \varepsilon,$$

в качестве параметров $\boldsymbol{\beta}$ возьмем, например, вектор $(\beta_0, \beta_1, \beta_2, \beta_3)^T = (1, -2, 6, 3)^T$. В нашей модели признаки \mathbf{x}_4 и \mathbf{x}_5 являются шумовыми. Результатом работы алгоритма является последовательность весов признаков, выбираемых на каждом шаге. В данном случае алгоритм сделает шесть шагов.

В таблице (1) представлены результаты работы алгоритма.

EXAMPLE

Т а б л и ц а 1. Результаты работы LALR

№	1	2	3	4	5	6
0	0	0	0	0.1969	0.2606	9.2868
1	0	0	-0.0250	-0.3142	-0.4733	-21.4689
2	0.7769	1.3359	1.4005	2.1215	2.5999	91.6048
3	0	0.3313	0.3615	0.6677	0.8713	36.5161
4	0	0	0	0	0	-8.2624
5	0	0	0	0	0.0513	1.6560

5. Заключение

В данной работе предложен и исследован новый алгоритм LALR, решающий задачу отбора признаков в модели логистической регрессии.

EXAMPLE

Список литературы

- [1] Smith H. Draper N. *Applied Regression Analysis*. Wiley, New York, 1966.
- [2] Friedman J. Hastie T., Tibshirani R. *The Elements of Statistical Learning: Data mining, Inference and Prediction*. Springer, New York, 2001.
- [3] Tibshirani R. Walther G. Hastie T., Taylor J. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.
- [4] Федоров В. В. Сухарев А. Г., Тимохов А. В. *Курс методов оптимизации*. ФИЗМАТЛИТ, Москва, 2005.

6. Приложение

Доказательство 1 (Леммы 1). Используя EXAMPLE разложим $\ell(\boldsymbol{\mu}_{\mathcal{A}_+})$ до первого члена,

$$\ell(\boldsymbol{\mu}_{\mathcal{A}_+}) = \text{EXAMPLE}. \quad (4)$$